BY CHRISTINE M. ANDERSON-COOK AND LU LU

Best Bang for Your Buck-Part 1

The size of experiments relative to design performance

WHEN DESIGNING AN experiment for a study, there are many choices to make, such as: what design factors to consider, which levels of the factors to use and which model to focus on. One aspect of design, however, is often left unquestioned: the size of the experiment.

When learning about design of experiments, problems are often posed as "select a design for a particular objective with N runs." It's tempting to consider the design size as a given constraint in the design-selection process.

If you think of learning through designed experiments as a sequential process, however, strategically planning for the use of resources at different stages of data collection can be beneficial: Saving experimental runs for later is advantageous if you can efficiently learn with less in the early stages. Alternatively, if you're too frugal in the early stages, you might not learn enough to proceed confidently with the next stages. Hence, choosing the right-sized experiment is important—not too large or too small, but with a thoughtful balance to maximize the knowledge gained given the available resources.

It can be a great advantage to think about the design size as flexible and include it as an aspect for comparisons. Sometimes you're asked to provide a small design that is too ambitious for the goals of the study. If you can show quantitatively how the suggested design size might be inadequate or lead to problems during analysis—and also offer a formal comparison to some alternatives of different (likely larger) sizes—you may have a better chance to ask for additional resources to deliver statistically sound and satisfying results.

14-run design example

We were recently approached by an engineering colleague who wanted us to suggest a 14-run design for a screening experiment involving seven factors in which the primary goal was to estimate the main effects model. There was a concern, however, that some two-way interactions or curvature for at least one factor might exist.

An easy option would be to create a computer generated D-optimal design with some statistical software that allows for a good estimation of the model parameters. In recent years, however, there has been discussion on the danger of an oversimplified decision for finding an optimal design based on only one criterion and the benefits of looking at multiple criteria when examining the appropriateness and desirability of designs for the goals of our experiment.[1-3]

In this case, we thought that given the constraints of the problem, a 14-run design might be ambitious to accomplish all that the engineer wanted. So we created four alternative designs (labeled 14r, 15rCR, 15DSD and 16r) in JMP[4] to present as potential solutions:

- 14r: A 14-run D-optimal design.
- 15rCR: A 15-run design (consisting of the 14-run design above with one center point added).
- 15DSD: A 15-run definitive screening design.[5]
- 16r: A 16-run D-optimal design (a non-regular design).[6]

As statisticians, our goal was not to provide a single answer, but to lead the discussion of alternatives so the engineer was informed to make a good decision to meet the study's needs.

In this column, we examine different criteria that should be balanced with cost when evaluating a design, and compare the four candidate designs' performance based on these criteria. Next month, we'll cover how to use these quantitative evaluations on multiple aspects of a good design to make a final decision and justify the choice.

Design comparisons

To compare the designs, we considered traditional design diagnostics, evaluated anticipated power for different-sized coefficients, looked at the aliasing structure of different terms in the model, considered the ability of designs to identify curvature and compared the anticipated prediction variance for new observations throughout the design space.

Understanding these different performance aspects in the context of the design size can provide insight into whether a design is well-suited for the goals and expected outcomes of an experiment. Many of the numerical and graphical summaries that follow are available in the statistical software JMP,[7] with a new functionality to compare multiple designs added in JMP 13.

A first category of comparison considers alphabetic optimality criteria,[8] which have been the traditional choice for single-number summaries to characterize the performance of a design. They focus on good estimation of the model parameters (D- and A-optimality) or good prediction of new observations (G- and I-optimality).

We initially considered the main-effects model because our primary interest is in estimating the main effects of all of the factors. The left side of Table 1 (p. 46)

shows the relative performance of the four designs compared to the 14r design with the originally requested design size. Values greater than one mean that the alternative design performs better for that criterion than 14r. We also included a cost-efficiency that compares the relative size of the designs. Clearly, because all of the other designs are larger than the 14r, their cost-efficiencies are less than one.

Looking across the D-, A-, G- and I-efficiencies, the general pattern is that larger designs perform better. This matches what would be expected: Collect more data and you will be able to learn more. The exception is the 15DSD, which does not perform as well for any alphabetic optimality criteria.

Comparison of several quantitative measures of the four designs / TABLE 1

Designs	Design diagnostics					Average correlation		
	D-eff.	A-eff.	G-eff.	I-eff.	Cost-eff.	Main × main	Main × inter.	Inter. × inter.
14r	1	1	1	1	1	0.059	0.231	0.164
15rCR	1.009	1.009	1.003	1.021	0.933	0.059	0.231	0.164
15DSD	0.841	0.793	0.791	0.844	0.933	0.167	0	0.276
16r	1.174	1.200	1.155	1.200	0.875	0	0.102	0.071

eff. = efficiency
inter. = interactions

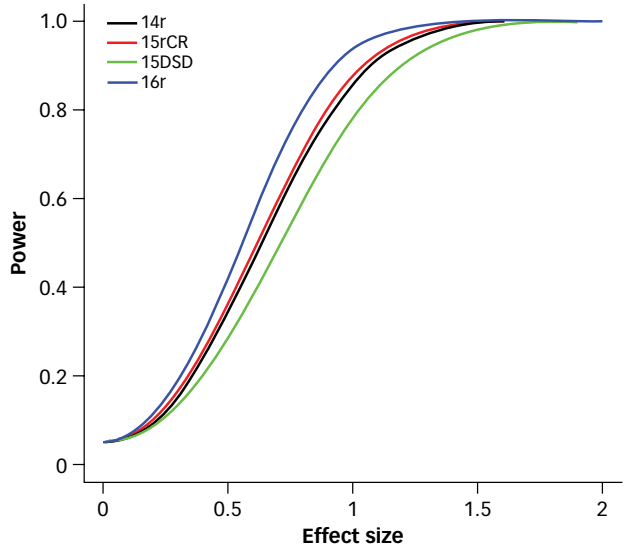
Evaluating the power of designs

Next, we considered the power of each design, which summarizes the ability to find a particular effect of a given size statistically significant during the analysis stage. Because all four designs are symmetric in how well they estimate all the main effects, looking at the performance for any individual main effect can be applied to all of the seven main effect terms in the model.

Figure 1 shows power curves for different sizes of the main effects for the four designs, assuming all have the same standard error (σ). The x-axis shows the effect size of the main effects, which is measured by its relative size to σ (that is, effect size equals one if the main effects equal σ) All designs start with 5% power (equal to the significance level) for zero main effects, and the power increases as the size of main effects increases until all designs reach 100% power for finding main effects of at least 1.5σ.

There is an obvious difference, however, in the power of the four designs for detecting small to moderate-sized effects (less than 1.5σ) in a general pattern as 16r>15rCR≈14r>15DSD. At an effective size of one, for example, the power values from top to bottom are 0.937 (16r), 0.877 (15rCR), 0.858 (14r) and 0.780 (15DSD). This means that for an effect of this size, there is a 93.7% chance that the 16r will find it statistically significant (with a p-value less than 0.05), while only a 78% chance that the 15DSD will find it significant.

Power summary / FIGURE 1



Note: This shows the estimated probability of finding any main effect of a given size statistically significant at the 5% level (that is, p-values less than 0.05) for the four designs.

Correlation considerations

Another consideration is how well the designs can estimate different potential terms in a larger model. For example, you might want to look for possible two-way interactions that are active and be able to identify likely candidates that influence the response without these effects being confounded with other effects in the model.

To understand the correlation's structure, you can examine a color map of the individual correlations. Figure 2 shows this plot that displays the absolute values of the pairwise correlations between all

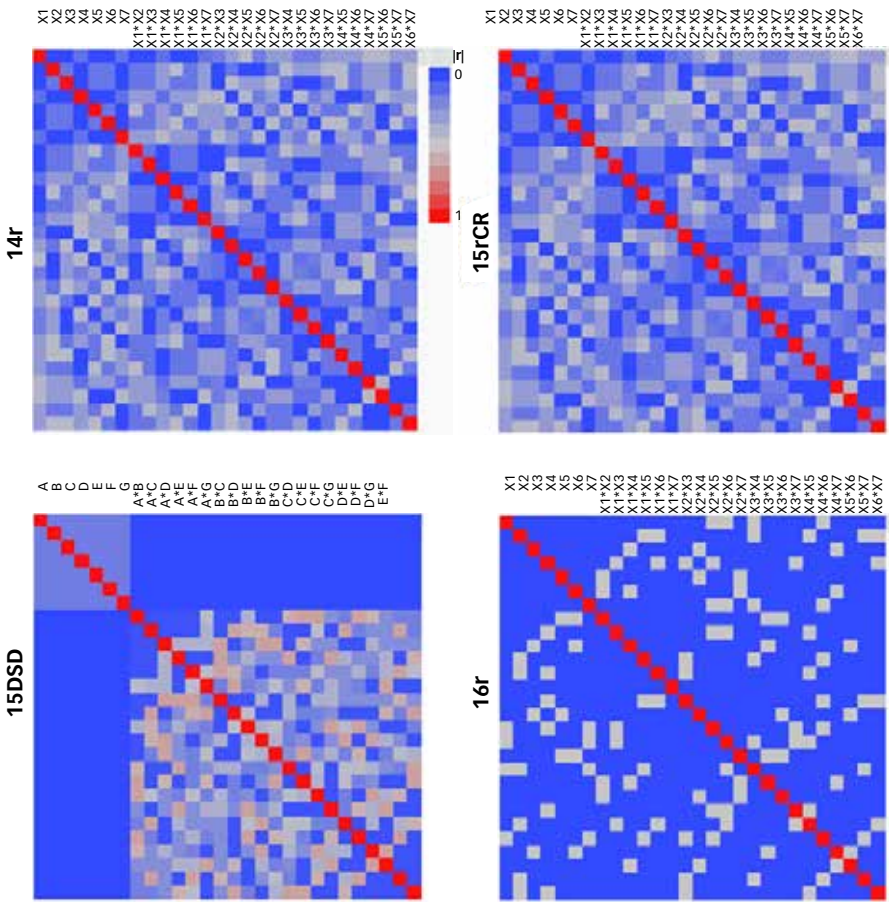
terms in a first-order model with all two-way interactions for the four designs. The ideal color is blue, which corresponds to small correlations. The darkest blue means the terms are uncorrelated, and these effects can be estimated separately from each other. White represents moderate correlations, while pink and red indicate pairs of effects in which there might be concern about being able to separate the contributions.

The top-left corner (7 x 7 block) shows the main x main correlations. The top right (7 x 21) and bottom left (21 x 7) rectangles show the correlations between main effects and two-way interactions, while the bottom-right corner (21 x 21 block) summarizes interaction x interaction pairs. Note that all designs have red diagonal elements because all terms have correlation one with themselves.

When you compare the four designs, overall impressions are that the 14r and 15rCR designs have small fractions of dark blue (uncorrelated pairs of effects). The 15DSD and 16r designs have a larger fraction of dark blue with the 15DSD having the entire main x interaction blocks being uncorrelated. The 16r design has some moderate white values, but the majority of pairs are uncorrelated, and it is the only design that has main effects completely unconfounded. The 15DSD has the largest absolute correlations for some pairs of terms with some pink shades in the interaction x interaction block.

To summarize over different groups of terms in the color map, the right side of Table 1 shows the average absolute correlation for between main effects (the top-left diagonal block), main effects with two-way interactions (the two off-diagonal blocks) and between two-way interactions (the bottom-right diagonal block). It should be clear that two similar averages in different designs could be achieved through different structures. You might have a design with a few large correlations and many uncorrelated pairs that could have a similar average to a design with all small correlations. Hence, the plot in Figure 2 can provide more details about how these averages were obtained that aren't possible to see from just Table 1. In terms of the average correlations, the 16r design is best for the main x main and interaction x interaction pairs, while the 15DSD is best for main x interaction pairs. The 16r design is quite appealing without pink or red squares and many blues. But this comes at the cost of using the largest design. The remaining three designs have more correlated pairs, with the 15DSD having some interaction x interaction pairs with moderately large correlations (shown in pink).

Correlation color map for 14r, 15rCR, 15DSD and 16r designs / FIGURE 2



Predicting new response values

Next, consider the precision of the designs to predict new observations throughout the design space for values of any factor between (-1 and +1). The fraction of design space (FDS) plot[9] shows a cumulative distribution of the prediction variances throughout the seven-dimensional design space. Note the FDS plot is an efficient way of understanding the precision of prediction across any dimension or shaped design space.

The ideal design has a relatively flat

curve (similar prediction for all locations in the design space) with small values. All of the designs have a best-prediction variance a bit smaller than 0.1, while the worst-prediction variance is for the 15DSD with a maximum value near 0.7. If you look at the median prediction variance (x-axis at 0.5), the four designs have the following values: 0.206 (16r) < 0.244 (15rCR) ≈ 0.244 (14r) < 0.286 (15DSD).

Clearly, to predict the response for a particular combination of factor levels, having as much precision as possible is beneficial. Not surprisingly, the largest design yields the most precision, and the FDS plot helps quantify the differences between choices.

Assessing curvature

Finally, consider the ability of the designs to assess curvature in the underlying response. Designs with only two levels (-1 and +1) for each factor (such as the 14r and 16r designs) are unable to make any determination of the presence or absence of curvature in the form of a quadratic term in the model.

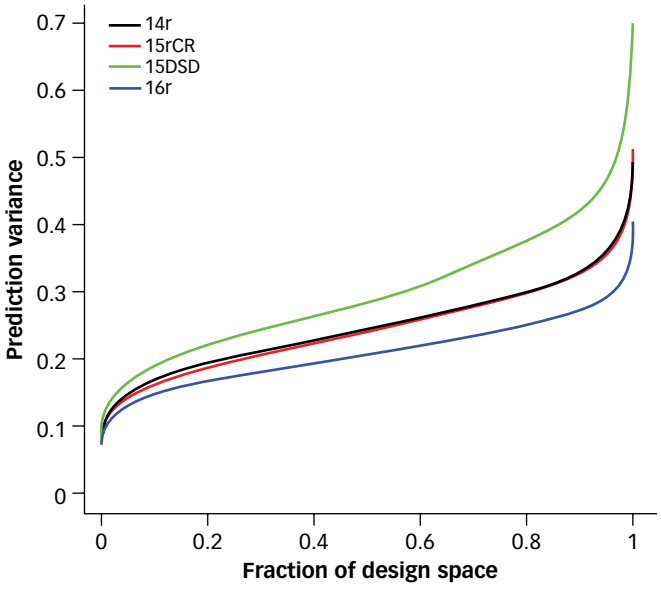
The 15rCR design has a single center run, which allows for an informal check of curvature of all of the factors simultaneously—namely, you could examine whether the value at the center run appears to match the estimated value of the response based on the chosen model. If it seems too different, you would suspect that at least one quadratic term should be added, but you would have no ability to decide which factors are associated with it unless more experimental data were collected. The only design that can evaluate all of the quadratic effects separately is the 15DSD. Hence, on this aspect, the definitive screening design provides a substantial advantage.

Understanding differences

Clearly, there are numerous trade-offs to consider when evaluating a design, and the four designs being compared have different strengths and weaknesses, as well as different associated costs. Comparing alternatives with quantitative summaries helps to understand the differences to make more-informed decisions.

Next month, we'll use what the engineer knows about the response and describe how to frame a compelling argument for a design size to match the experimental goals. QP

Fraction of design space plots / FIGURE 3



Note: This shows the quantiles of the prediction variance across the seven-dimensional design region for factor level values in [-1 and +1] for the four designs.

REFERENCES

1. Lu Lu, Christine M. Anderson-Cook and Timothy J. Robinson, "Optimization of Designed Experiments Based on Multiple Criteria Utilizing a Pareto Frontier," *Technometrics*, Vol. 53, No. 4, 2011, pp. 353-365.
2. Lu Lu and Christine M. Anderson-Cook, "Rethinking the Optimal Response Surface Design for a First-Order Model With Two-Factor Interactions, When Protecting Against Curvature," *Quality Engineering,* Vol. 24, No. 3, 2012, pp. 404-422.
3. Lu Lu, Christine M. Anderson-Cook and Dennis Lin, "Optimal Designed Experiments Using a Pareto Front Search for Focused Preference of Multiple Objectives," *Computational Statistics and Data Analysis,* Vol. 71, 2014, pp. 1,178-1,192.
4. JMP, version 13, SAS Institute Inc., 2016.
5. Bradley Jones and Christopher J. Nachtsheim, "A Class of Three-Level Designs for Definitive Screening in the Presence of Second-Order Effects," *Journal of Quality Technology,* Vol. 43, No. 1, 2011, pp. 1-15.
6. Lu Lu, Mark E. Johnson and Christine M. Anderson-Cook, "Selecting a Best Two Level 16-Run Screening Design From the Catalog of Non-Isomorphic Regular and Non-Regular Designs for Six to Eight Factors," *Quality Engineering,* Vol. 26, No. 3, 2014, pp. 269-284.
7. JMP, version 13, see reference 4.
8. Raymond H. Myers, Douglas C. Montgomery and Christine M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments,* fourth edition, Wiley, 2006, pp. 126-128.
9. Alyaa Zahran, Christine M. Anderson-Cook and Raymond H. Myers, "Fraction of Design Space to Assess the Prediction Capability of Response Surface Designs," *Journal of Quality Technology,* Vol. 35, No. 4, 2003, pp. 377-386.

CHRISTINE M. ANDERSON-COOK is a research scientist in the Statistical Sciences Group at Los Alamos National Laboratory in Los Alamos, NM. She earned a doctorate in statistics from the University of Waterloo in Ontario. Anderson-Cook is a fellow of ASQ and the American Statistical Association.

LU LU is an assistant professor in the department of mathematics and statistics at the University of South Florida in Tampa. She was a postdoctoral research associate in the statistical sciences group at Los Alamos National Laboratory. She earned a doctorate in statistics from Iowa State University in Ames, IA.